LM-KBC 2023: 2nd Challenge on Knowledge Base Construction from Pre-trained Language Models

Jan-Christoph Kalo[1], Sneha Singhania[2], Simon Razniewski[3] and Jeff Z. Pan[4]

[1]*University of Amsterdam, The Netherlands*
[2]*Max Planck Institute for Informatics, Germany*
[3]*Bosch Center for AI, Germany*
[4]*The University of Edinburgh*

Abstract

Large language models (LLMs) like chatGPT [1] have advanced a range of semantic tasks and are being ubiquitously used for knowledge extraction. Although several works have explored this ability by crafting prompts with in-context or instruction learning, the viability of complete and precise knowledge base construction from LMs is still in its nascent form. In the 2nd edition of this challenge, we invited participants to extract disambiguated knowledge triples from LMs for a given set of subjects and relations. In crucial difference to existing probing benchmarks like LAMA [2], we made no simplifying assumptions on relation cardinalities, i.e., a subject-entity can stand in relation with zero, one, or many object-entities. Furthermore, submissions needed to go beyond just ranking predicted surface strings, and materialize disambiguated entities in the output, which were evaluated using established KB metrics of precision, recall, and F1-score. The challenge had two tracks: (1) a small model track, where models with < 1 billion parameters could be probed, and (2) an open track, where participants could use any LM of their choice. We received seven submissions, two for track 1 and five for track 2. We present the contributions and insights of the submitted peer-reviewed submissions and lay out the possible paths for future work. All the details related to the challenge can be found on our website at https://lm-kbc.github.io/challenge2023/.

Keywords

Large Language Models, Knowledge Base Construction, Information Extraction, Prompting

1. About the Challenge

Background Large-scale Language Models (LLMs) like BERT [3], LLaMA [4], and chatGPT [1] have been optimized to predict missing parts of textual inputs or complete sentences and have significantly improved performances on various NLP tasks such as question answering and machine translation. Lately, these LLMs have been recognized for their potential ability to generate structured knowledge directly from their parameters. This is a promising development since existing knowledge bases (KBs) like Wikidata [5], DBpedia [6], YAGO [7], and ConceptNet [8] are crucial components of the Semantic Web ecosystem but are inherently limited and incomplete due to manual or (semi) automatic construction.

KBC-LM: Knowledge Base Construction from Pre-trained Language Models workshop at ISWC 2023

j.c.kalo@uva.nl (J. Kalo); ssinghan@mpi-inf.mpg.de (S. Singhania); Simon.Razniewski@de.bosch.com (S. Razniewski); j.z.pan@ed.ac.uk (J. Z. Pan)
https://people.mpi-inf.mpg.de/~ssinghan/ (S. Singhania); http://simonrazniewski.com/ (S. Razniewski)
0000-0000 (S. Singhania); 0000-0000 (J. Z. Pan)

CEUR Workshop Proceedings (CEUR-WS.org)

The groundbreaking LAMA paper by Petroni et al. [2] demonstrated promising results on knowledge extraction from pre-trained language models. Over the years, there have been several advancements and criticisms in follow-up research, investigating the possibility of using LLMs for KB construction [9, 10, 11, 12, 13]. New datasets, but also a variety of new techniques to either probe LLMs for factual knowledge or to construct KBs directly from the language model, have been proposed. To establish a competitive venue for evaluating this promising research, we introduced the Knowledge Bases from Pre-trained Language Models (LM-KBC) challenge at the International Semantic Web Conference 2022 [14].

This paper describes the 2nd edition of the LM-KBC challenge at ISWC 2023, with the following changes compared to the previous year: (1) we added an entity disambiguation component to overcome problems associated with entity aliases and also to enable participants to probe LLMs directly for entity identity, (2) a new and larger dataset, comprising diverse relationships incorporating popular, long-tail entities, and literal values, and (3) a small track for LMs with up to 1B parameters, and an open track for any LM choice.

Task Description Given an input pair consisting of a subject-entity s and a relation r, the objective is to generate accurate object-entities $[o_1, o_2, o_3, ..., o_k]$ by probing the language model.

In contrast to last year, we added an entity disambiguation component to the task. While language models are working on natural language, KBs usually work with abstract identifiers to prevent ambiguities among labels. Hence, a KB can distinguish between synonym entities. For example, in Wikidata Athens, the capital of Greece is identified by Q1524, while the city of Athens in Ohio, USA, is identified by Q755420. Hence, in this edition, we ask the participants to predict Wikidata identifiers, instead of just names.

For instance, Table 1 illustrates GPT-3 performance when probed with a sample prompt containing a subject-entity and relation pair, and a set of few-shot examples. The model returns the predictions in the format provided in the few-shot examples, in this case, a list.

Similar to last year, the challenge had two tracks:

- Track 1: **Small model track**, where LMs with up to 1 billion parameters were allowed;
- Track 2: **Open track**, where LMs of arbitrary size, including retrieval-augmented models, could be used.

LM-KBC'23 Dataset Compared to last year, where we had only 12 relations, this edition presents 21 relations, comprising of diverse set of subjects and a complete list of ground-truth objects. Each relation has a maximum of 100 unique subject entities in all data splits. Table 2 provides more details. The object entities could be person, organization, country, count entities, or even "none". The ground truth consists of the exact identifiers from Wikidata [5].

Evaluation Each test instance prediction is evaluated using precision, recall, and f1-score. In the following, we show the metrics calculation. Let P be the prediction list of object entities for a test subject-entity, and GT be its corresponding ground-truth list of object entities, then the metrics are:

Input Sample	Prompt	LM Prediction	Ground Truth
(Japan, CountryOfficialLanguage)	Cologne, CityLocatedAtRiver, [Rhine] Hexadecane, CompoundHasParts, [carbon, hydrogen] Antoine Griezmann, FootballerPlaysPosition, [forward] Japan, CountryOfficialLanguage,	[Japanese]	[Q5287]
(Italy, CountryBordersCountry)	State of Palestine, CountryBordersCountry, [Q801] Paraguay, CountryBordersCountry, [Q155, Q414, Q750] Lithuania, CountryBordersCountry, [Q34, Q36, Q159, Q184, Q211] Italy, CountryBordersCountry,	[Q142]	[Q142]

Table 1
Example predictions on few-shot prompts from GPT-3 on the LM-KBC 2023 dataset.

Relation	|Train|	|Val|	|Test|	Train	Val	Test
BandHasMember	100	100	100	[2, 15]	[2, 16]	[2, 16]
CityLocatedAtRiver	100	100	100	[1, 9]	[1, 5]	[1, 9]
CompanyHasParentOrganisation	100	100	100	[0, 5]	[0, 3]	[0, 5]
CompoundHasParts	66	66	66	[2, 6]	[2, 5]	[2, 6]
CountryBordersCountry	63	63	63	[1, 17]	[1, 10]	[1, 17]
CountryHasOfficialLanguage	65	65	65	[1, 16]	[1, 11]	[1, 16]
CountryHasStates	46	46	46	[1, 20]	[1, 20]	[1, 20]
FootballerPlaysPosition	100	100	100	[1, 2]	[1, 3]	[1, 2]
PersonCauseOfDeath	100	100	100	[0, 1]	[0, 3]	[0, 1]
PersonHasAutobiography	100	100	100	[1, 4]	[1, 4]	[1, 4]
PersonHasEmployer	100	100	100	[1, 6]	[1, 13]	[1, 6]
PersonHasNobelPrize	100	100	100	[0, 1]	[0, 2]	[0, 1]
PersonHasNumberOfChildren	100	100	100	[1, 1]	[1, 2]	[1, 1]
PersonHasPlaceOfDeath	100	100	100	[0, 1]	[0, 1]	[0, 1]
PersonHasProfession	100	100	100	[1, 11]	[1, 12]	[1, 11]
PersonHasSpouse	100	100	100	[1, 3]	[1, 3]	[1, 3]
PersonPlaysInstrument	100	100	100	[1, 8]	[1, 8]	[1, 8]
PersonSpeaksLanguage	100	100	100	[1, 10]	[1, 4]	[1, 10]
RiverBasinsCountry	100	100	100	[1, 9]	[1, 5]	[1, 9]
SeriesHasNumberOfEpisodes	100	100	100	[1, 2]	[1, 1]	[1, 2]
StateBordersState	100	100	100	[1, 16]	[1, 12]	[1, 16]

Table 2
Dataset Characteristics. For each of the 21 relations, the number of unique subject entities in the train, dev, and test are given in the GitHub repo. The minimum and maximum number of object entities for each relation is given below. If the minimum value is 0, then the subject entity can have zero valid object entities for that relation.

$$\text{Precision} = \frac{P \cap GT}{|P|} \qquad \text{Recall} = \frac{P \cap GT}{|GT|} \qquad f1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

When P is empty, and GT is not, precision $= 1$ and recall $= 0$, leading to $f1 = 0$. On the other hand, when GT is empty, recall $= 1$ but precision $= 1$ only when P is empty, else precision $= 0$, leading to either 1 or 0 $f1$-score. Scores were macro-averaged across subjects and relations, and the final macro-$f1$-score ranked systems.

2. Systems

Participants submitted their predictions on CodaLab at https://codalab.lisn.upsaclay.fr/competitions/14777 to get scores on the private test dataset. The leaderboard can be seen in Table 4Below, we explain the baselines and provide insights from the seven submissions.

2.1. Baselines

We provide several baselines:

- Standard prompt for HuggingFace models with Wikidata default disambiguation: These baselines can be instantiated with various HuggingFace models (e.g., BERT, OPT), generate entity surface forms, and use the Wikidata entity disambiguation API to generate IDs.
- Few-shot GPT-3 directly predicting IDs: This baseline uses a few samples to instruct GPT-3 to directly predict Wikidata IDs.
- Few-shot GPT-3 w/ NED: Like above, but predicting surface forms disambiguated via Wikidata's default disambiguation.

Baseline Performance is shown in Table 3.

Method	Avg. Precision	Avg. Recall	Avg. F1-score
GPT-3 NED (Curie model)	0.308	0.210	0.218
GPT-3 IDs directly (Curie model)	0.126	0.060	0.061
BERT	0.368	0.161	0.142

Table 3
Baseline Performance

2.2. Track 1

Winner: Expanding the Vocabulary of BERT for Knowledge Base Construction
Dong Yang, Xu Wang and Remzi Celebi

The submission utilizes the BERT language model to improve knowledge base construction, specifically addressing the challenge of multi-token object extraction. A novel approach called "Token Recode" is introduced to expand the model's vocabulary while retaining semantic context. This method shows a noticeable improvement in f1 scores, confirming its effectiveness. The study also employs task-specific pre-training and category-based filtering to enhance performance further, achieving promising results even with a lightweight BERT model. The code for this system is available at https://github.com/MaastrichtU-IDS/LMKBC-2023.

Broadening BERT vocabulary for Knowledge Graph Construction using Wikipedia2Vec
Debanjali Biswas, Stephan Linzbach, Dimitar Dimitrov, Hajira Jabeen and Stefan Dietze

The paper proposes an interesting novel method for predicting multi-token entities with the BERT model by using Wikipedia2Vec [15] embeddings. The resulting model is trained with the prompt tuning technique OptiPrompt [16]. While this is an exciting new idea, the performance of the resulting system is similar to the baseline model BERT. Further experiments would be needed to understand these results further. The code for this system is available at https://github.com/debanjali05/LM-KBC2023-GESIS.

2.3. Track 2

Winner: Using Large Language Models for Knowledge Engineering (LLMKE): A Case Study on Wikidata
Bohui Zhang, Ioannis Reklos, Nitisha Jain, Albert Meroño Peñuela and Elena Simperl

The paper outlines a two-step pipeline for KBC: knowledge probing and entity mapping. In the probing step, GPT-3.5 Turbo and GPT-4 are utilized. Three types of settings are tested: question prompting, triple completion prompting, and context-enriched prompting. Few-shot learning techniques are used across all settings to improve result formatting. In the entity mapping step, the MediaWiki Action API is used to obtain candidate Wikidata entities for object strings. Three methods are employed for final disambiguation: case-based, keyword-based, and LM-based. The code for this system is available at https://github.com/bohuizhang/LLMKE.

Limits of Zero-shot Probing on Object Prediction
Shrestha Ghosh

The paper introduces a system called "Minimal Probe", which mainly emphasizes two areas: prompt design and answer post-processing. For prompt design, the study constructs prompts consisting of a task description, optional demonstrations, and the task itself. It shows that introducing a well-framed task description can improve model performance substantially over a baseline. In answer post-processing, the paper employs manually designed cleaning steps to ensure the answers align with the desired format. The code for this system is available at https://github.com/ghoshs/LM-KBC2023.

Knowledge-centric Prompt Composition for Knowledge Base Construction from Pre-trained Language Models
Xue Li, Anthony Hughes, Majlinda Llugiqi, Fina Polat, Paul Groth and Fajar J. Ekaputra

The authors introduce a pipeline for constructing knowledge bases using large language models, specifically GPT-3.5 and GPT-4. The research explores various configurations involving in-context learning, employing an example selector and knowledge-enriched prompts for better contextual relevance. Findings indicate that rule-based example selectors, which consider cardinality per relation, significantly improve performance. Additionally, augmenting entities and relations with extra properties sourced from GPT-4 further enhances the system's effectiveness. The code for this system is available at https://github.com/effyli/lm-kbc/.

Enhancing Knowledge Base Construction from Pre-trained Language Models using Prompt Ensembles
Fabian Biester, Daniel Del Gaudio, and Mohamed Abdelaal

The paper centers on the idea of "prompt ensembles" for improving knowledge base construction from language models. The researchers initially used baseline prompts with ChatGPT and then only the top-performing ones. Then, a few shot learning approach on LLAMA2 [4] with 70b parameters is performed. As a last step, a fact-checking step is performed. The code for this system is available at https://github.com/asdfthefourth/lmkbc.

LLM2KB: Constructing Knowledge Bases using instruction tuned context aware Large Language Models
Anmol Nayak and Hari Prasad Timmapathini

This paper uses instruction-fine-tuned LLAMA2 [4] and StableBeluga [17] models with LoRa [18] together with dense passage retrieval to extend the prompt. The authors prepared a Wikipedia corpus with textual data about the subject entities of interest and put them into the FAISS index for usage with DPR [19]. Then, they fine-tune two LLMs with three instruction prompts on the training dataset. Wikipedia paragraphs extend these instruction prompts via DPR. A similar process is performed at inference time. However, an additional entity disambiguation step is performed, where the Wikidata API baseline disambiguation method is used to retrieve candidate entities that are then sent to an LLM to perform disambiguation via prompting. The code for this system is available at https://github.com/anmoln94/Team_LLM2KB_LM-KBC-2023.

3. Discussions

The second edition of our *LM-KBC* challenge received encouraging uptake, with seven teams going past the finish line and submitting both code and system descriptions. Table 4 presents the final leaderboard of our challenge.

3.1. Takeaway

The main findings across all the submissions are:

1. **Larger models beat the smaller models by a large margin.** We observe that the submissions on track 2, which mainly consist of relation-specific prompts for probing the GPT-4 model, have a considerably higher performance than track 1 submissions.

System	Track	Precision	Recall	F1-score
Zhang et al.	2	71.5 %	72.6 %	70.1 %
Li et al.	2	71.3 %	69 %	67.4 %
Biester et al.	2	66.4 %	65.8 %	62.5 %
Nayak and Timmapathini	2	73.2%	60 %	61.2 %
Ghosh	2	64.6 %	62.6 %	60.9 %
GPT-3 Baseline	2	30.8 %	21 %	21.8 %
GPT-3 Baseline (IDs directly)	2	12.6 %	6 %	6.1 %
Yang et al.	1	39.5 %	39.3 %	32.3 %
Biswas et al.	1	13.3 %	23.2 %	14 %
BERT Baseline	1	36.8 %	16.1 %	14.2 %

Table 4
Challenge leaderboard showcasing the final scores on the test dataset.

2. **Possible saturation using GPT-4 probing.** The top submissions on the leaderboard, which correspond to track 2, have a similar performance range across all three metrics. Since the prompts are manually crafted and relation-specific, the similarity in performance could be due to similar prompt formulations. However, post-processing the prediction list leads to the highest performance.

3. **Easy vs hard relations.** Across submissions, we notice that the 21 relations could be categorized as (1) top-5 easy relations: [PersonHasNoblePrize, CompoundHasParts, CountryHasOfficialLanguage, RiverBasinsCountry, PersonCauseOfDeath], and (2) top-5 hard relations: [PersonHasEmployer, PersonHasAutobiography, PersonHasProfession, StateBordersState, BandHasMember]. This insight could help us better curate the next edition dataset by considering the object list cardinality and entity popularity.

3.2. Possible Extensions

Deciding on the challenge complexity required navigating a trade-off between ease of access and realism. Several avenues for extension are:

1. **Only small scale models:** Given the easy API access to GPT-4 model variants and the associated monetary cost, it might be interesting only to host the challenge with a small-scale model track. This might lead to innovative solutions focusing on retrieval augmentation, model compression, and knowledge transfer.

2. **Temporal object list:** Real-world knowledge bases with factual information keep evolving with new information. A special track incorporating the time component into the object entities to study the consistency and stability of LMs for knowledge base construction could be helpful.

3. **Other metrics:** Our evaluation focused on macro-averaged f1-scores, which give equal weight to precision and recall. It could be interesting to explore other trade-offs; as for KBs, precision is often way more critical than recall. Also, as subjects with no objects

dominate many domains (e.g., very few people hold political offices), a higher presence or more weight on no-object subjects might be interesting.

4. Acknowledgments

We thank Huawei for sponsoring a 1000 Euro prize for the winning systems. We also thank the semantic web challenge chairs, Valentina Ivanova and Wen Zhang, for helping us host a successful second edition of our challenge. Finally, we thank the participating teams for their enthusiasm and contributions.

References

[1] OpenAI, GPT-4 technical report, 2023. URL: https://doi.org/10.48550/arXiv.2303.08774. doi:10.48550/arXiv.2303.08774. arXiv:2303.08774.

[2] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, A. Miller, Language models as knowledge bases?, in: Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 2463–2473. URL: https://aclanthology.org/D19-1250. doi:10.18653/v1/D19-1250.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, NAACL (2019). URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv (2023).

[5] D. Vrandečić, M. Krötzsch, Wikidata: A free collaborative knowledgebase, Commun. ACM 57 (2014) 78–85. URL: https://doi.org/10.1145/2629489. doi:10.1145/2629489.

[6] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: The semantic web, 2007, pp. 722–735.

[7] F. M. Suchanek, G. Kasneci, G. Weikum, Yago: A core of semantic knowledge, in: Proceedings of the 16th International Conference on World Wide Web, WWW '07, 2007, p. 697–706. URL: https://doi.org/10.1145/1242572.1242667. doi:10.1145/1242572.1242667.

[8] H. Liu, P. Singh, Commonsense reasoning in and over natural language, in: Knowledge-Based Intelligent Information and Engineering Systems, 2004, pp. 293–306.

[9] C. Wang, X. Liu, D. Song, Language models are open knowledge graphs, arXiv preprint arXiv:2010.11967 (2020).

[10] S. Hao, B. Tan, K. Tang, H. Zhang, E. P. Xing, Z. Hu, Bertnet: Harvesting knowledge graphs from pretrained language models, arXiv preprint arXiv:2206.14268 (2022).

[11] B. Veseli, S. Singhania, S. Razniewski, G. Weikum, Evaluating language models for knowledge base completion, in: European Semantic Web Conference, 2023, pp. 227–243.

[12] B. Veseli, S. Razniewski, J.-C. Kalo, G. Weikum, Evaluating the knowledge base completion potential of gpt, Findings of EMNLP (2023).

[13] J. Z. Pan, S. Razniewski, J.-C. Kalo, S. Singhania, J. Chen, S. Dietze, H. Jabeen, J. Omeliya-

nenko, W. Zhang, M. Lissandrini, et al., Large language models and knowledge graphs: Opportunities and challenges, TGDK (to appear) (2023).

[14] S. Singhania, T.-P. Nguyen, S. Razniewski, LM-KBC: Knowledge base construction from pre-trained language models, Semantic Web challenge (2022).

[15] I. Yamada, A. Asai, J. Sakuma, H. Shindo, H. Takeda, Y. Takefuji, Y. Matsumoto, Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia, in: Empirical Methods in Natural Language Processing: System Demonstrations, Online, 2020, pp. 23–30. URL: https://aclanthology.org/2020.emnlp-demos.4. doi:10.18653/v1/2020.emnlp-demos.4.

[16] Z. Zhong, D. Friedman, D. Chen, Factual probing is [MASK]: Learning vs. learning to recall, in: North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 5017–5033. URL: https://aclanthology.org/2021.naacl-main.398. doi:10.18653/v1/2021.naacl-main.398.

[17] S. AI, Stable beluga, 2023. URL: https://stability.ai/blog/stable-beluga-large-instruction-fine-tuned-models.

[18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. arXiv:2106.09685.

[19] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6769–6781. URL: https://aclanthology.org/2020.emnlp-main.550. doi:10.18653/v1/2020.emnlp-main.550.